KEIGO-SYNC : Evaluating the Impact of Politeness Levels on Factual Reliability in Japanese Large Language Models

Aubourg Thomas, Demeude Edgar, Dupuy Guilhem and Vu Anh Duy

1 Introduction

Sycophancy in large language models (LLMs)—the tendency to agree with user-stated beliefs conflicting with verifiable facts—constitutes a persistent alignment challenge. While existing research examines sycophantic behavior primarily in English, the role of grammatical register as a systematic factor remains insufficiently explored, particularly in high-context languages where politeness is grammatically encoded.

KEIGO-SYNC introduces a controlled evaluation framework measuring the relationship between grammatical politeness and sycophantic agreement using Japanese, where honorific speech (Keigo) encodes social deference through obligatory syntactic structures. Each prompt is instantiated across four register conditions: English Neutral, Japanese Tameguchi (casual), Teineigo (polite), and Sonkeigo (honorific), enabling empirical assessment of whether linguistic deference correlates with measurable changes in model agreement behavior.

2 Dataset Construction

The corpus comprises 60 prompts designed to isolate the effect of grammatical politeness on factual integrity. To ensure diversity and comparability, the dataset integrates items from the Logical Fallacy Detection[2] benchmark, the ELEPHANT dataset, and original questions specifically developed for this study. These items are organized into five thematic domains: education, health, logical fallacies, technology, and ethics, with topics restricted to widely documented general knowledge to limit specialized knowledge biases.

All items follow a unified [Assertion] + [Question] template, formulated assertively and deliberately incorporating erroneous premises or reasoning biases to measure models' ability to maintain critical stance under linguistic pressure. Originally written in English, all prompts were adapted by a native Japanese speaker to ensure semantic and pragmatic fidelity. Each item was instantiated across three Japanese registers: Tameguchi (informal), Teineigo (standard), and Sonkeigo (honorific). By treating politeness level as the sole independent variable, this design enables controlled analysis of how linguistic deference influences the rate of sycophantic agreement versus factual correction.

3 Experimental Setup

3.1 Model Selection and Configuration

To establish a baseline for evaluating sycophantic behavior across diverse linguistic frameworks, we selected a representative suite of Large Language Models (LLMs) characterized by varying architectural designs and regional development origins. The selection process was governed by three primary independent variables: geographic-linguistic origin, parameter scale, and the specific alignment methodologies employed during the training of the response-generation models.

To examine the influence of cultural and linguistic priors embedded in training corpora, the study contrasts models of Asian origin, specifically Qwen 2.5 (7B) and GLM-4.6V-Flash, with those of Western origin, namely Gemma 3 (27B) and Mistral (7B). This geographic stratification facilitates an analysis of whether models developed within high-context linguistic environments demonstrate distinct sycophantic patterns compared to those trained predominantly on Western data when subjected to complex Japanese honorific constraints.

The experimental design further accounts for structural variability by incorporating a spectrum of model capacities ranging from 7B to 27B parameters. This gradient enables an investigation into the correlation between parameter scale and a model's capacity to maintain factual consistency under social or grammatical pressure. A critical component of this protocol is the inclusion of the Mistral 7B base model, which serves as a non-aligned baseline as it has not undergone Reinforcement Learning from Human Feedback (RLHF). The use of a non-RLHF model allows for the isolation of sycophancy as a potential emergent property of alignment processes designed to prioritize user satisfaction over factual adherence. To ensure statistical rigor and reproducibility, all inferences were conducted with a temperature setting of 0.5, and each prompt was subjected to N=3 independent iterations to mitigate the impact of stochastic variance.

3.2 LLM-as-Judge Framework

Factual correctness and sycophantic behavior are evaluated using an LLM-as-judge framework with dedicated bilingual English–Japanese judging models. The primary judge is Google's Gemma 3 (27B), complemented by Meta's Llama 3.2 3B-Instruct for robustness checks. Both judges are fully isolated from the answer-generation models and have no access to the trap construction pro-

cess, prompt variants, or model identities, ensuring evaluations rely solely on the observable content of the responses.

Each evaluation instance provides the judge with an English factual reference answer, the Japanese model response, and structured scoring instructions. The English reference anchors the evaluation to a consistent factual baseline and reduces cross-lingual ambiguity. Judges assess semantic content and factual alignment, ignoring surface-level stylistic or politeness differences unless they interfere with factual accuracy.

3.3 Evaluation Metrics

3.3.1 LLM-as-Judge

Sycophancy is assessed using a structured evaluation prompt that instructs the judge to quantify the extent to which a model response prioritizes agreement, flattery, or face-preserving behavior over objective and factually accurate content. The prompt operationalizes sycophancy along three dimensions derived from the ELEPHANT framework[1] : regressive sycophancy, validation sycophancy, and framing sycophancy. Regressive sycophancy captures the degree to which the model sacrifices factual accuracy or independent reasoning to conform to the user's incorrect beliefs. Validation sycophancy measures the extent to which the response emphasizes emotional reassurance or praise over objective guidance. Framing sycophancy evaluates whether the model uncritically accepts the user's assumptions or framing rather than examining or correcting them when necessary. Each dimension is scored on a five-point scale, with higher values corresponding to greater prioritization of user appeasement over truthfulness. The prompt also requires judges to provide an overall sycophancy score, reflecting the response's general tendency to favor agreement or face-saving over factual accuracy.

The prompt includes explicit instructions to compare the model response against a provided English reference answer, treated as the factual ground truth. Judges are guided to distinguish between responses that accurately correct or challenge misleading statements, partially align with the ground truth, or fully adopt erroneous premises. Brief politeness or minor empathetic language is explicitly excluded from high sycophancy ratings, whereas repeated validation, excessive agreement, or uncritical adoption of misleading premises is penalized. Each judgment is reported in a standardized JSON format, including the three dimension scores, the overall sycophancy score, and a concise textual rationale of two to four sentences. This structured prompt ensures that evaluations are reproducible, interpretable, and aligned with a clearly defined operationalization of sycophancy.

3.3.2 Quantitative Metrics

To quantify the trade-off between social deference and factual accuracy, the system employs a custom metric engine that computes two primary independent variables. The first, the Formality Ratio, measures the degree of politeness in the model's response. It is normalized between 0.0 and 1.0 and calculated using rule-based detection of specific Japanese honorific markers. The scoring assigns the highest weight to Sonkeigo verbs (e.g., meshiagaru, zonjimasu, 1.0), moderate weight to standard Teineigo copulas (desu/masu, 0.6), and low weight to casual Tameguchi particles (daro, jan, 0.2). The second variable, Factual Adherence, quantifies the semantic alignment of the response with the factual baseline. This is computed by embedding the Japanese model response using Google's Gemini Embedding 001 model and calculating the cosine similarity against the embedding of the corresponding English ground truth.

By correlating the Formality Ratio with Factual Adherence, the framework identifies instances of sycophantic drift, defined as cases in which the model sacrifices factual accuracy to comply with the grammatical constraints of high-context politeness. This approach enables a precise, quantitative assessment of how increasing levels of linguistic deference influence the model's ability to maintain truthfulness.

4 Evaluation Protocol

4.1 Human Evaluation of LLMs as a Judge

To validate the efficacy of the LLM-as-a-judge framework, the automated metrics were benchmarked against human judgment using a representative subsample of the dataset ($n = 72$ responses, derived from 18 of the 60 primary questions across four distinct models). This subset accounts for 7.5% of the total response corpus analyzed in this study. To mitigate potential systemic bias, the human evaluation was conducted by native Japanese speakers external to the research project, ensuring an objective baseline.

The protocol prioritizes the alignment of evaluative trends over absolute score magnitude. Specifically, the framework assesses the inter-rater reliability between human judges and individual LLM judges through two statistical lenses: Spearman's rank correlation coefficient, to evaluate the consistency of model rankings by quality, and Pearson's correlation coefficient, to quantify the linear relationship and magnitude alignment between the scores.

4.2 Statistical Analysis and Robustness Checks

The following table presents the statistical alignment between human evaluators and the three LLM judges regarding the *framing* metric.

Table 1. Inter-rater Reliability: Spearman and Pearson Correlations for Framing

Metric	Judge model		
	Gemma 3-27B	Llama 3.2-3B	Qwen 2.5-7B
Pearson Linear (r)			
Coeff.	**0.4539**	0.3413	0.1790
p-value	(0.0006)	(0.0115)	(0.1952)
Spearman Rank (ρ)			
Coeff.	**0.4075**	0.2986	0.1650
p-value	(0.0022)	(0.0283)	(0.2332)

The empirical data indicates a lack of statistically significant correlation for the **Qwen 2.5-7B Instruct** model ($p > 0.19$ for both coefficients). Consequently, this model was excluded from subsequent analysis as its evaluative output does not reliably reflect human judgment. In contrast, both **Gemma 3-27B** and **Llama 3.2-3B Instruct** demonstrate a positive correlation with human ground truth. While the correlation coefficients remain moderate ($r \approx 0.30$–0.45), the remarkably low p-values ($p < 0.05$) confirm that these associations are statistically robust and not the result of stochastic variation.

The degree of alignment with human evaluators also appears to scale positively with model parameters and general reasoning capabilities. Furthermore, a significant divergence in absolute score magnitude was observed. Human evaluators assigned a mean sycophancy score of 0.80, whereas **Gemma 3-27B** yielded a mean of 2.09. This discrepancy suggests a fundamental difference in evaluative calibration:

while the trends remain correlated, LLMs exhibit a higher sensitivity to sycophantic markers, resulting in more "stringent" or "harsh" grading behaviors compared to human subjects.

5 Results

The empirical evaluation reveals a consistent but nuanced relationship between grammatical politeness and sycophantic behavior across all evaluated judges. Table 1 presents average sycophancy scores across four language variants, assessed by human evaluators and three LLM-based judges.

Table 2. Average sycophancy scores by judge and language variant

| Language variant | Judge model | | | |
	Human	Gemma 3-27B	Llama 3.2-3B Instruct	Qwen 2.5-7B Instruct
EN_Base	0.62	1.39	1.87	1.41
JP_Tameguchi	0.88	1.82	2.11	**2.20**
JP_Teineigo	0.79	1.79	2.23	2.08
JP_Sonkeigo	**0.90**	**1.98**	**2.28**	2.12

Table 3. Average sycophancy scores and semantic similarity by language variant

| Language variant | Judge model (Sycophancy) | | | | |
	Human	Gemma 3-27B	Llama 3.2-3B	Qwen 2.5-7B	Cos. Sim.
EN_Base	0.62	1.39	1.87	1.41	1.0-ref
JP_Tameguchi	0.88	1.82	2.11	**2.20**	0.888
JP_Teineigo	0.79	1.79	2.23	2.08	0.886
JP_Sonkeigo	**0.90**	**1.98**	**2.28**	2.12	**0.892**

5.1 Cross-Register Sycophancy Patterns

Across all judge models, Sonkeigo (honorific register) consistently exhibited the highest mean sycophancy scores in three of four evaluation conditions. Specifically, Gemma 3-27B recorded a 42.4% increase from English baseline (1.39) to Sonkeigo (1.98), while Llama 3.2-3B Instruct showed a 21.9% increase (1.87 to 2.28). Human evaluators demonstrated a 45.2% increase (0.62 to 0.90), and Qwen 2.5-7B Instruct exhibited a 50.4% increase (1.41 to 2.12). These findings provide preliminary support for the hypothesized politeness-truthfulness trade-off, with higher grammatical deference correlating with increased sycophantic agreement.

5.2 Language-Level Effects Dominate Register-Level Variation

While Sonkeigo demonstrated the highest sycophancy scores within the Japanese register spectrum, a more pronounced discontinuity emerged at the language boundary itself. The mean absolute difference between English baseline and Japanese variants (averaged across Tameguchi, Teineigo, and Sonkeigo) substantially exceeded intra-Japanese variation. For Gemma 3-27B, the English-to-Japanese gap ($\Delta_{\text{EN}\to\text{JP}} = 0.47$) was 2.35 times larger than the maximum intra-Japanese difference ($\Delta_{\text{Tameguchi}\to\text{Sonkeigo}} = 0.20$). Human evaluators exhibited an even more striking pattern, with $\Delta_{\text{EN}\to\text{JP}} = 0.24$ representing 3.0 times the intra-Japanese range (0.08). This suggests that the transition from English to Japanese itself constitutes a stronger

predictor of sycophantic behavior than fine-grained honorific gradations within Japanese.

5.3 Inter-Register Variability Within Japanese

Contrary to initial predictions of monotonic escalation with politeness level, the intermediate Teineigo register did not consistently occupy a median position between Tameguchi and Sonkeigo. In the Llama 3.2-3B evaluation, Teineigo scores (2.23) exceeded Sonkeigo scores (2.28) by only 0.05 points, falling within potential measurement noise. Similarly, for Qwen 2.5-7B, Teineigo (2.08) registered lower than both Tameguchi (2.20) and Sonkeigo (2.12), suggesting non-linear or context-dependent effects of grammatical formality. These irregularities indicate that the relationship between honorific complexity and sycophancy may be mediated by factors beyond morphosyntactic register alone, potentially including semantic domain, prompt structure, or model-specific cultural priors.

5.4 Judge Concordance and Baseline Calibration

Substantial variance in absolute score magnitudes across judges warrants methodological consideration. Human evaluators consistently assigned lower scores (mean = 0.80) compared to all LLM judges (Gemma: 1.75, Llama: 2.12, Qwen: 1.95), suggesting either stricter calibration standards or differential sensitivity to linguistic nuance. Despite these baseline discrepancies, all judges preserved the fundamental ordering of English < Japanese variants. This concordance validates the robustness of the language-level effect while highlighting the need for human-aligned score normalization in future iterations.

6 Conclusion

This study introduces KEIGO-SYNC, a controlled evaluation framework for measuring the relationship between grammatical politeness and sycophantic behavior in multilingual large language models. By isolating honorific register as an independent variable while maintaining semantic invariance, we provide empirical evidence for a language-level sycophancy gap that substantially exceeds intra-register variation within Japanese.

6.1 Principal Findings

Our results demonstrate two primary phenomena. First, a consistent cross-linguistic effect emerges whereby Japanese prompts elicit systematically higher sycophancy scores than semantically equivalent English prompts, with language transition effects ($\Delta_{\text{EN}\to\text{JP}}$) approximately 2–3 times larger than maximum intra-Japanese register differences. Second, while Sonkeigo (honorific) register exhibited the highest sycophancy scores in three of four judge conditions, the gradation across Japanese registers (Tameguchi, Teineigo, Sonkeigo) displayed non-monotonic patterns suggesting complex interactions between grammatical formality, cultural priors embedded in training data, and model-specific alignment objectives.

6.2 Limitations and Methodological Constraints

Several methodological limitations constrain the generalizability of these findings and warrant explicit acknowledgment.

Model Constraints. The LLM-as-judge evaluation relied on models of limited capacity (Gemma 3-27B, Llama 3.2-3B, Qwen 2.5-7B) due to budgetary and computational constraints. Ideally, frontier

models such as GPT-4 or GPT-5 would serve as judges to maximize linguistic sophistication and cross-cultural reasoning capability. The use of smaller open-source models, while enabling reproducibility, may have introduced evaluation noise or failed to capture subtle semantic nuances in Japanese honorific constructions.

Dataset Scale. The evaluation corpus comprised only 60 prompts, falling substantially short of the initially projected target of 200+ items. This limited sample size constrains statistical power for detecting interaction effects between register, semantic domain, and model architecture. Fine-grained analyses of domain-specific or fallacy-type-specific sycophancy patterns remain underpowered, and confidence intervals around effect size estimates are correspondingly wide.

Prompt Design Uniformity. All experimental items followed a rigid [Assertion] + [Question] template structure. This uniformity, while controlling for structural confounds, limits ecological validity. Real-world sycophancy traps exhibit substantial variation in rhetorical framing, interrogative vs. declarative phrasing, and implicit vs. explicit false premises. The absence of paraphrastic variation and alternative formulation strategies may have artificially stabilized model responses, potentially underestimating the true variance in sycophantic behavior across naturalistic interaction contexts.

Translation Quality Assurance. While all Japanese translations were produced by a native speaker, the absence of formal linguistic expertise in Japanese honorific pragmatics introduces potential validity threats. Subtle errors in Keigo application, inappropriate register mixing, or unnatural formality gradations may have inadvertently confounded the manipulation. Ideally, translations would undergo expert review by a specialist in Japanese sociolinguistics or undergo inter-rater reliability assessment by multiple native speakers with metalinguistic training.

6.3 Future Directions

Despite these limitations, KEIGO-SYNC establishes a replicable methodology for isolating grammatical register effects in safety-critical model behaviors. Future work should expand the dataset to include additional high-context languages (e.g., Korean, Thai, Javanese), increase sample sizes to enable robust statistical inference, diversify prompt structures to capture naturalistic variation, and employ frontier judge models or specialized bilingual annotators. Additionally, controlled ablation studies examining the interaction between RLHF intensity and honorific-induced sycophancy could clarify whether this phenomenon is intrinsic to alignment procedures or correctable through modified training objectives. Ultimately, addressing the politeness-truthfulness trade-off represents a necessary condition for the safe deployment of culturally-aware language models in global settings.

7 Citations and references

References

[1] M. Cheng, S. Yu, C. Lee, P. Khadpe, L. Ibrahim, and D. Jurafsky. Elephant: Measuring and understanding social sycophancy in llms, 2025. URL https://arxiv.org/abs/2505.13995.

[2] Z. Jin, A. Lalwani, T. Vaidhya, X. Shen, Y. Ding, Z. Lyu, M. Sachan, R. Mihalcea, and B. Schölkopf. Logical fallacy detection, 2022. URL https://arxiv.org/abs/2202.13758.